心理学报 2019, Vol. 51, No. 10, 1116−1127
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2019.01116

基于事件相关电位(ERPs)和机器学习的考试焦虑诊断*

章文佩[1,2] 沈群伦[3] 宋锦涛[1] 周仁来[1]

([1]南京大学心理系，南京 210023) ([2]安徽工业大学工商管理系，马鞍山 243032)
([3]中国科学院数学与系统科学研究院，北京 100190)

摘 要 考试焦虑对个体的身心具有严重危害。传统诊断考试焦虑的方法容易受到个体主观态度的影响，从而影响对个体考试焦虑的发现与及早干预。为了克服传统主观问卷对考试焦虑群体诊断的不足，本研究提出脑电神经数据结合机器学习的客观综合诊断方法评估个体的考试焦虑水平。研究采用情绪 Stroop 范式，结合脑电技术测量个体对考试焦虑者的注意抑制功能，机器学习基于此前提，提取 P1, P2, N2, P3 和 LPP 五种事件相关电位(ERP)成分，以卷积神经网络(CNN)为主采用 7 种常见的机器学习算法对个体考试焦虑程度进行进一步的诊断。结果表明 CNN 对考试焦虑诊断的准确率达 86.5%，F1-score 为 0.911，显著高于其他 6 种常见算法。因此采用 CNN 对脑电信号进行深度学习得出的诊断模型能够有效地对个体的考试焦虑程度进行诊断。

关键词 机器学习；考试焦虑；情绪 Stroop; ERPs
分类号 R395

1 前言

在中国，考试是评价个人能力的一种主要手段。作为一种评价性事件，个体对考试的认知会影响个体在面对此类事件时的焦虑程度(Schutz, Davis, & Schwanenflugel, 2002)。当个体非常重视考试结果并因此将考试视为一种威胁，会出现焦虑的症状(Lotz & Sparfeldt, 2017)。高度考试焦虑在各级学生中的比例都超过 20% (陈睿，刘潇楠，周仁来，2011)，在一些地区的初中生中这一比例甚至高达 35% (陈祉妍, 2002)。

考试焦虑对个体的身心健康具有严重危害。首先，在面临重要考试的时候，考试焦虑者会体会到一种强烈的担忧和情绪反应，时时刻刻在担心考试的失败，他人的评价以及考试结果的排名(Zeidner & Matthews, 2005)。不仅是心慌、紧张等情绪体验，考试焦虑者还会表现出一系列生理的、行为的反应，如出现与植物性神经活动失调相关的身体反应症状(如心跳加快、出冷汗、呼吸急促、颤抖等)，并且由此影响个体的生理健康，影响内分泌，降低免疫功能，增加感染性疾病、胃部不适及睡眠障碍等躯体疾病(闫慧，卢莉，2014)。此外，高度的考试焦虑往往与抑郁情绪相联系(陈睿等，2011)，而抑郁症发病得越早，越可能影响至终生，且复发率和自杀率越高(王玮文，谢希，邵枫，2008)。因此，对考试焦虑进行早期的准确诊断十分必要。

目前，国内外对考试焦虑的研究中所采用的诊断技术主要采用主观测评，具体分为问卷调查法和访谈法。然而，想要仅通过主观测评技术达到对考试焦虑进行早期识别并准确诊断其程度并不容易。具体限制在于：(1)真实性：当被试由于某些原因想要隐藏自己真实的情况时，采用访谈或者问卷等主

收稿日期: 2018-10-29
* 中央高校基本科研业务费专项资金(14370303)、江苏省普通高校学术学位研究生科研创新计划项目(KYZZ16_0010)和安徽省高校人文科学研究项目(SK2017A0084)资助。
章文佩与沈群伦为共同第一作者
通信作者：周仁来, E-mail: rlzhou@nju.edu.cn

观测评的方法可能会降低评估的准确性(风笑天, 2003), 如学生不希望自己被老师或者家长知道自己的考试焦虑, 则会选择更为积极的主观表述, 从而影响评估结果。(2)诱导性: 主观评估的过程可能激发个体的负性情绪, 如在填答问卷中看到"考试", "挂科"等字面负性信息, 或者访谈中提及有关考试的负性经历, 个体都有可能会诱发出相关的负性情绪(Diegomantecón, 2015), 从而可能会加重或者影响个体作答时的情绪状态, 从而影响评估结果。

为了降低这些限制, 结合客观技术的综合诊断必不可少。而精确敏感的客观技术指标需要以考试焦虑的病理模式为基础。考试焦虑者并不总是处在一种不适应的状态中, 考试焦虑症状的出现具有情境性和特异性(Lowe et al., 2008), 当没有或者面对非重要考试相关事件时, 考试焦虑者并没有明显的心理生理不适症状, 只会表现出一定的焦虑症状, 但当重要考试相关事件出现时, 考试焦虑者的焦虑水平会急速上升, 伴随着明显的心理生理反应, 并进一步影响个体的认知水平(Lotz & Sparfeldt, 2017; Mok & Chan, 2016)。这表明, 考试焦虑的症状是随着考试焦虑者对考试事件的认知而变化的, 即考试焦虑者越将考试事件视作是一种威胁(即对考试的认知越不合理), 越能够激发他们的不适症状(Mochcovitch, da Rocha Freire, Garcia, & Nardi, 2014)。因此, 考试焦虑者的认知模式是对考试焦虑进行诊断的重要基础。

脑电技术(Electroencephalography, EEG) 可以有效反映个体对特定刺激的情绪状态变化、注意及背后的认知模式(Edwards, Burt, & Lipp, 2010)。脑电是人脑活动时产生的自发电位, 具有较高的时间分辨率和敏感性 (Luck, Woodman, & Vogel, 2000), 其中, 事件相关电位(event-related potentials, ERPs)则是大脑对特定类型刺激的电位反应, 可以反映个体对特定事件的认知模式。考试焦虑者的重要认知特点为将考试视为一种威胁, 因此考试相关威胁信息出现时, 高考试焦虑者更容易将注意资源放在考试相关威胁信息上(即注意偏向), 并持续加工这些信息, 对当前需要进行的任务产生干扰(Kalanthroff, Henik, Derakshan, & Usher, 2016; Putwain, Langdale, Woods, & Nicholson, 2011)。具体表现为当考试相关(威胁)信息出现时, 高考试焦虑者在 ERP 的重要成分上有显著的波幅变化(增加或降低)。

情绪 Stroop 范式能够很好地反映高考试焦虑个体对考试威胁信息的认知特点(van Bockstaele et al., 2014; Verhaak, Smeenk, van Minnen, & Kraaimaat, 2004)。在情绪 Stroop 任务中, 每次给被试呈现一个词语刺激, 同时包含目标维度(颜色)和干扰维度(词义), 要求被试只专注目标维度(即判断词的颜色)而忽略干扰维度(即词义), 词义分为考试相关威胁词(如: 挂科)和中性词(如: 街道), 任务通过比较威胁词和中性词条件下的 ERP 成分变化推断个体对威胁信息的注意特点(Dennis & Chen, 2009; Gu et al., 2011)。由于高考试焦虑者对考试相关威胁信息存在注意偏向, 因此, 当威胁词出现时, 相比于低考试焦虑者, 高考试焦虑者会在一些有重要意义的 ERP 成分上表现出波幅的显著变化, 可能表现为相对早期的、感觉的, 与自动化加工关的成分(如 P1, P2, N2 等成分) (Kanske & Kotz, 2012; Wabnitz, Martens, & Neuner, 2016), 及相对晚期的、认知的, 与自主加工和情绪活动相关的成分(如 P3, LPP 等成分; Albert, López-Martín, & Carretié, 2010; Raz, Dan, Arad, & Zysberg, 2013)的波幅显著增强。

采用 ERPs 技术对考试焦虑程度评估可以有效降低问卷法的限制: (1)真实性: ERPs 中的特定成分反映的是个体对特定刺激的自动化反应, 不易于自主控制, 具有高度的客观性(Righi, Mecacci, & Viggiano, 2009)。(2)诱导性: ERP 任务中呈现的刺激材料往往时间较短, 个体没有充分的时间对其进行加工, 从而对个体(Morel, George, Foucher, Chammat, & Dubal, 2014; Tillman & Wiens, 2011)的情绪和认知影响较低。然而 ERPs 技术也存在自身的局限性: (1)个体差异性: 不同个体之间的脑电幅值可能差异很大, 很难找到具有代表性的有效特征, 使得在使用脑电对不同群体进行分类诊断的精确性受到影响(Boshra, Ruiter, Reilly, & Connolly, 2016; 王艳娜, 孙丙宇, 2017)。(2)干扰性: 由于脑电指标十分敏感, 因此很容易受到外界干扰信号或者内部其他类型认知的干扰(Cecotti et al., 2011)。因此单独使用 ERPs 技术进行分类诊断的准确率无法保证。

为了减少 ERPs 技术的限制对考试焦虑诊断性的影响, 增加诊断的稳定性与准确性, 我们在脑电技术的基础上进一步采用机器学习技术。机器学习特别是深度学习是一种强力的分类模型, 已经在图像识别(Krizhevsky, Sutskever, & Hinton, 2012), 自然语言处理(Kumar et al., 2016), 文本分类(Yang et al., 2016)等任务中取得很好的结果。我们主要采用卷积神经网络(Convolutional neural network, CNN)

这种深度学习算法对脑电类型数据进行模型的建立。在适用性方面, CNN 是一种基于普通神经网络的推广算法, 特别善于捕捉数据的局部特征。脑电数据虽然存在个体差异性和干扰性的局限, 但是也存在相对稳定性, 即在头皮上相邻电极点之间的点位变化具有很大的相关性, 结合分析能够提高准确性。而 CNN 可以组合分析相邻电极点之间的脑电数据, 通过下采样的方式来减小数据矩阵的大小, 有效减少数据的位移、扰动和一些小的变化对数据稳定性和准确性的影响, 因此 CNN 对脑电数据具有高度适用性(Lu, Jiang, & Liu, 2017; Seijdel, Ramakrishnan, Losch, & Scholte, 2016)。在具体操作方面, 为了处理一些复杂的任务, 在传统的分类模型中, 往往需要对数据进行很复杂的特征提取, 然后将得到的特征放入分类模型中进行处理。而 CNN 是一种端对端的算法, 即只需要将经过简单预处理的数据作为模型的输入, CNN 会自动学习特征, 并且利用习得的特征进行分类。此外, 同传统机器学习方法相比, CNN 在这一类有空间结构的数据上表现远超传统机器学习方法(Lee, 2015; Fotin, Haldankar, & Periaswamy, 2016), 并且已被验证确实能够提取出高层次的有用的信息(Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015), 同时神经网络的结构能够保证它可以实现对任何一个从输入向量到输出向量的连续映射函数的逼近(Hornik, 1991)。所以我们认为 CNN 能在 ERPs 数据上取得良好的结果。

因此, 本文主要关注考试焦虑的程度评估与诊断问题, 采用卷积神经网络(CNN)对高、低考试焦虑者在情绪 Stroop 中的 ERP 脑电信号进行分类模型的建立, 并进一步使用该模型对被试的考试焦虑进行诊断, 试图探究更为客观、准确的考试焦虑诊断方法。

2 数据采集与预处理

2.1 被试招募

本研究通过海报及网络招募的方式招募了 82 名被试。被试(年龄为 18~26 岁; 皆为右利手)根据考试焦虑量表(Sarason, 1978)得分以及两位专家的综合评估被分至高考试焦虑组(TAS 分数: 27.85 ± 4.78, 人数为 57 人, 男性 25 人, 年龄: 21.27 ± 1.89 岁)和低考试焦虑组(TAS 分数: 8.65 ± 2.76, 人数为 25 人, 男性 12 人, 年龄: 21.35 ± 2.96 岁)。该实验已经通过伦理委员会的审查, 所有被试在实验前已经签署知情同意书, 均为自愿参加实验, 在实验之后也获得相应的报酬(40 元)。

2.2 考试焦虑量表(TAS)

考试焦虑量表是由美国临床心理学家 Irwin G. Sarason 于 1978 年编制完成的(Sarason, 1978)。TAS 量表共 37 题, 每个问题要求作是或否的二择一回答, “是”记 1 分, “否”记 0 分, 通过计算总分对考试焦虑程度进行评估, 总分范围为 0~37, 得分越高说明考试焦虑的程度越高, TAS 得分≥20 为高考试焦虑者, TAS 得分≤12 为低考试焦虑者(Newman, 1996; Wang, 2001)。量表的重测信度为 0.61, 同质性系数为 0.64。量表的结构效度采用与考试焦虑测验(TAI)的相关测得, TAS 总量表分和 TAI 的担心(worry)分量表的相关为 0.48; 和 TAI 的情绪性(emotionality)分量表的相关为 0.60 (王才康, 2001)。

2.3 情绪 Stroop 任务

情绪 Stroop 任务设计与前人设计类似(Thomas, Johnstone, & Gonsalvez, 2007), 要求被试忽略词义, 只判断词的颜色。在材料上: (1)词义分为两种条件: 考试相关威胁词(如“试卷”, “分数”)和中性词(如“花园”, 鞋子)。词汇的选取是通过评定的方法: 请 40 位被试(不参加此次实验)根据威胁度和相关度筛选出考试相关威胁词与中性词各 15 个, 并根据使用频率进行匹配。评定结果为考试相关威胁词的威胁度($t(38) = 30.19$, $p < 0.001$)与相关度($t(38) = 38.166$, $p < 0.001$)都显著高于中性词, 且两类词在使用频率上没有显著差异($t(38) = 1.436$, $p = 0.162$)。(2)词色分为两种条件: 红色和蓝色。在操作上, 任务包括两部分(1)练习部分: 包含 6 次实验试次, 但是每次呈现的都是中性词, 具体设置与实验部分(见后文)类似, 且练习部分中出现的词都没有出现在实验部分中。此外, 每个试次中在被试进行反应之后程序都呈现“正确”或“错误”的反馈(实验部分不呈现反馈); (2)实验部分: 包含 120 次试次(每个词汇随机出现 4 次, 2 次为红色, 2 次为蓝色)。每个试次都以计算机屏幕中央呈现注视点“+”开始, 该注视点停留在屏幕上 200 ms, 之后屏幕呈现空白并持续一定时间(在 800 至 1200 ms 之间随机), 随后一个目标词将出现在白色背景下。每个试次在以下两种情况下结束: (a)被试完成反应(按下按钮选择词汇的颜色), 或者(b)在 2000 ms 内未进行反应。试次间会出现空白屏幕并持续一定时间(在 1000 至 1200 ms 之间随机)。

2.4 ERP 信号采集

本研究采用 NeuroScan 公司的 64 导放大器采

集 EEG 信号。采集时采用左侧乳突作为参考电极。水平眼电分别置于双眼外眼睑处，垂直眼电分别置于左眼上下 2.5 cm 处。全头电阻始终保持在 5 kΩ 以下。EEG 信号的采集采用直流电(DC)模式，分辨率为 1000 Hz。

2.5　数据预处理

脑电信号的离线处理采用 Curry 7.0.8 软件。EEG 信号通过双侧乳突进行转参考，进行 0~30 Hz 的滤波，并对垂直眼电与质量不佳的信号进行校正或删除。ERP 成分信息通过叠加被试分别在两种条件下的 EEG 信号得出：EEG 信号以每次刺激前 200 ms 至刺激后 1000 ms (共计 1200 ms)进行分段叠加，采用刺激前 200 ms 的数据作为基线对 ERP 波形进行校正。在具体分析的 ERP 成分上，本研究根据前人文献(Donaldson, Ait Oumeziane, Hélie, & Foti, 2016; Felmingham, Stewart, Kemp, & Carr, 2016)并结合本研究的结果提取出 5 个具有含义的 ERP 成分：P1 (120~170 ms), P2 (210~260 ms), N2 (240~290 ms), P3 (320~370 ms)和 LPP (450~600 ms)，在每个 ERP 成分的时间段内取峰值作为此成分的数据值。最终，对于每一个被试，我们采集有 64 个电极点信号，每个电极点包含威胁词，中性词两种条件，每种条件含有 5 种 ERP 成分的峰值数据，即一个被试有 64×2×5 = 640 个数据。为了确定这 5 种成分的选择是否具有代表性，我们对分别对 5 种成分在 Fz, FCz, Cz, CPz 和 Pz 五个电极点上的 ERP 波幅进行 2(组别高考试焦虑，低考试焦虑) × 2(条件：考试焦虑威胁词，中性词)的重复测量方差分析，从而判断这 5 种成分的选取是否能有效区分高、低考试焦虑者。

在神经网络任务中，我们一般会对数据进行归一化或者正则化处理，这样可以使模型尽快的收敛，由于这些数据的绝对值都小于 15，我们直接将数据除以 15，使它们的取值在(−1, 1)之间。

2.6　多折交叉验证

为了对每种机器学习算法进行更为客观的比较，我们采取 k 折交叉验证的方式，即：将样本均匀地分为互斥的 k 份，保证每一份的样本个数相同。一共进行 k 次训练，每次训练选其中 k−1 份作为训练集，剩下一份作为测试集，最终的指标为 k 次训练之后得到的模型在测试集上指标的平均值(见图 1)。一种基于经验的 k 值确定方式为 k≈log (n) (Jung, 2018), n 为样本量的大小。这里 log (n) = log (82) ≈4.4，因此我们向上取整取 k = 5，使用 5 折交叉验证。

3　卷积神经网络(Convolutional neural network, CNN)

3.1　卷积层

卷积操作是卷积神经网络的核心操作，通过它模型得以提取数据的不同特征，模型也是通过这一步在数据中学习到了卷积核的参数。卷积的操作如图 2，具体公式为：

$$g_{ij}(z_{i-1}) = \sum_{k=0}^{C_i-1} C_{h_{ijk}} z_{i-1}^{k}$$

这里 $g_{ij}: R^{m_{i-1}\times n_{i-1}\times c_{i-1}} \to R^{m_i\times n_i\times c_i}$ 的一个映射，m_i, n_i, c_i 分别表示第 i 层网络中输出的数据矩阵的长、宽、通道数。$C_g a$ 表示对图片 a 使用卷积核 g 进行卷积，z_i^j 表示第 i 层网络输出数据矩阵的第 j 个通道。

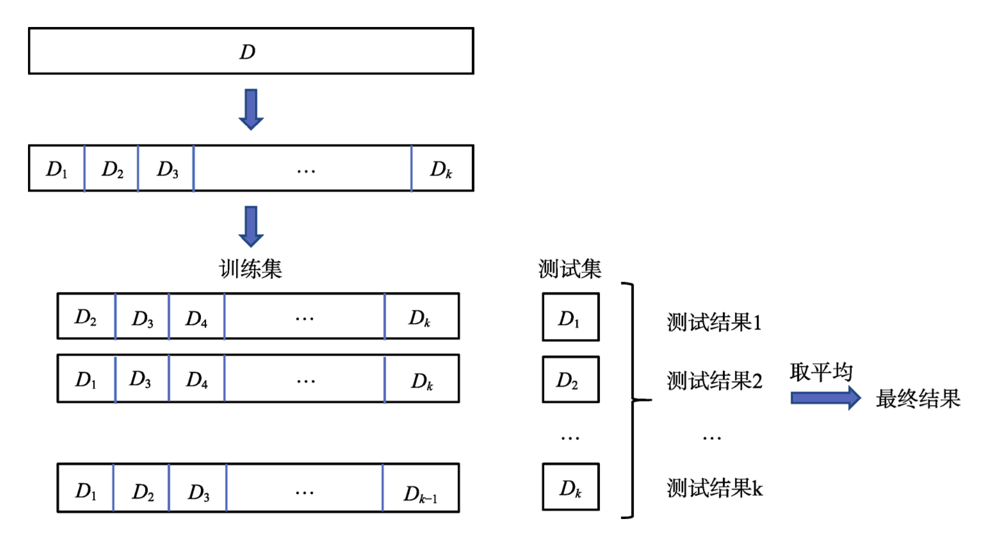


图 1　图中的 D 表示原始数据集，$D_1, D_2, \cdots D_k$ 表示将 D 分成的 k 个相同大小的子集

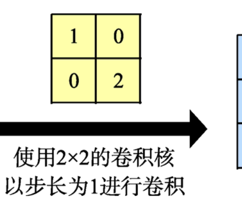


图 2 卷积操作的计算展示

注：这里的卷积是不进行补全的卷积，即卷积运算之后数据矩阵会相应变小，同时也有一种补全的卷积操作，即在原数据矩阵周围添 0，使得卷积之后得到的数据矩阵大小不变。

卷积操作是通过卷积核(红色矩阵)在数据矩阵(蓝色矩阵)上进行滑动，将对应的元素进行相乘相加得到的新元素作为输出矩阵的对应元素。这里输出数据矩阵的长和宽皆为：4−2+1 = 3。黄色矩阵的第一个元素是由 1×1+2×0+5×0+6×2 ＝ 13 得到，由于我们的步长是 1，那么将红色矩阵向右滑动一格，黄色矩阵的第二个元素由 2×1+0×3+6×0+1×2 = 4 得到，其他元素以此类推。使用多个卷积核就可以得到多个不同的输出，以此得到输入数据的多个不同特征，卷积核中的元素是所要训练的参数，可以通过反向传播的方式进行训练(LeCun & Bengio, 1995)。

3.2 池化层

池化是卷积神经网络中常用的一种操作，它通过降低矩阵长和宽的大小，降低了数据矩阵的分辨率，但是也进一步压缩并提取了原数据的特征，并且减少了网络计算的复杂度。图 3 是一个最大池化

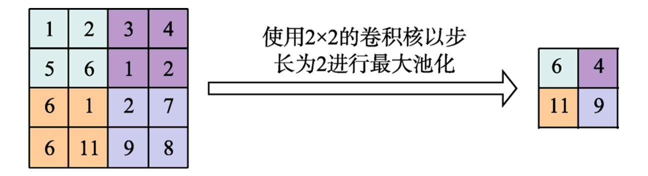


图 3 最大池化的计算展示

注：图中表示的是一个 4×4 的矩阵上使用一个 2×2 的窗口以步长为 2 进行最大池化，其原理就是取出每个 2×2 窗口中的的最大元素作为输出矩阵中对应元素的值。

操作的直观展示，不难看出，如果数据矩阵中的部分数据发生一些轻微变化，最大池化还是能输出一样的结果，这也是模型对于数据的偏移和旋转有较好的鲁棒性的原因。

3.3 网络架构

本研究使用的卷积神经网络架构见图 4。本文中使用的 CNN 的输入数据维度是 64×2×5，其中 64 代表 64 个不同位置的电极点，2 代表任务条件(即威胁词和中性词下不同的脑电信息)，5 代表 5 种 ERP 成分(即 P1，P2，N2，P3 和 LPP 成分)。将原始数据输入到卷积层 Conv1 做卷积计算，即用一个较小的卷积核(也叫卷积矩阵)在数据矩阵上根据给定的步长(这里步长为 1)进行滑动，将对应位置的元素进行相乘求和。在如图中输入数据矩阵为 64×2×5 的情况下，用 16 个 5×5 的卷积核来进行卷积操作，每一个卷积核都进行卷积操作就得到 16 个 64×2 的矩阵(这里我们使用补全的卷积方式，于是数据矩阵的大小并不发生改变)，这 16 个矩阵分别代表 16 种原数据的不同特征，在深度学习中我们称为通道数。可以看出卷积是一种局部操作，通过一定大小的卷积核作用于局部数据区域来提取局部信息，这里卷积核的大小是事先给定的，里面的参数由模型学习而来，这些特性使得 CNN 的参数可以共享，减少了参数个数，并且在数据发生平移变换的时候，模型仍能捕捉到相似的特征。为了满足不同任务的需要，近年来，许多不同的卷积核如空洞卷积也被提出(Yu & Koltun, 2015)。

卷积层一般会跟着一个下采样操作，又叫池化层，即通过一个小矩阵在数据矩阵上滑动，只提取小矩阵中的最大数据(最大池化)或平均数据(平均池化)，从一个较大数据矩阵压缩到一个较小的矩阵用作下一层的输入，这一层没有参数需要学习，通过池化运算可以减少分辨率，降低数据对噪音的

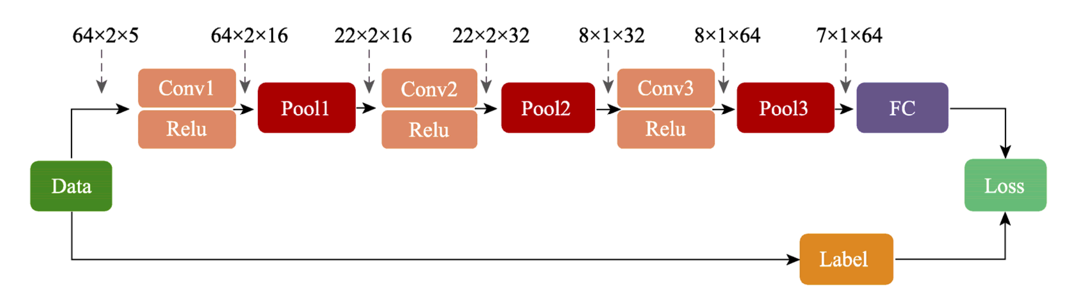


图 4 本研究使用的卷积神经网络架构图

注：每一层的具体参数见表 1。横线上的数据表示这一层的输入数据的维度，也即上一层输出数据的维度。Conv 代表卷积操作，Pool 代表池化操作，relu 代表在卷积操作之后的非线性激活方法。

敏感程度。在这里 Pool1 层做的就是池化操作，通过一个 4×1 的矩阵在上一卷积层的输出矩阵中以步长为 3 滑动，使该输出矩阵变为 22×2×16 的大小(22 = [65/4]+1, []表示向下取整)。

每次池化之后我们都会对输出的数据矩阵做非线性激活，这一手段使得模型能够拟合任意的数据流形，常用的非线性激活函数是 relu 函数。即：

$$relu(x)=max(0,x)$$

在很多实验中验证了这是一个非常有效的激活函数，并且具有生物学意义，这种非线性的激活函数使得卷积神经网络可以逼近任意数据分布，使得网络有了非常强大的拟合能力，同时也有防止梯度消失的作用(Nair & Hinton, 2010)。

在后面的几层中我们继续做了卷积和池化的操作，使得数据矩阵的长和宽越来越小，而通道数越来越多，即学习得到的特征数越来越多，这就是卷积神经网络的特征提取的过程。

传统的 CNN 在倒数几层架构中会将数据矩阵展平成一个向量，再加入几层全连接网络，也就是 FC 层，最后一层再用 softmax 分类器进行输出。而最近的研究指出全连接网络会非常容易导致过拟合，取消中间的全连接层，而全用卷积层代替，并且加入平均池化也即 Pool3，可以大大提高模型的泛化性(Lin, Chen, & Yan, 2013)，因此这里我们采用的也是这种架构。

最后在经过 FC 层之后能得到一个预测的类别即被试的考试焦虑或者非考试焦虑，将预测的结果与已知的实际类别进行比较并计算两者之间的误差，使用优化算法 Adam 来优化模型中每一层的参数从而减少误差，使预测的正确率不断上升，本文所使用的卷积神经网络在交叉验证下的正确率达到了 86.6%。本文中的卷积神经网络的代码框架是 TensorFlow，在 python 上进行了实现，并使用了 GPU 加速，显卡配置是 2 块 Quadro P500。

为了便于说明每一层的结构，我们将卷积和池化分为两层来介绍，这里我们建立了一个 7 层的卷积神经网络(表 1)，通过卷积操作来提取特征，通过池化来压缩数据的分辨率，最后采用平均池化提取全局特征，这一操作可以增加模型的泛化能力。

4　其他机器学习方法

本文还使用了其他机器学习的分类方法：逻辑回归(Logistic Regression), K 近邻(KNN)，支持向量机(SVM)，随机森林(Random Forest)，人工神经网络(ANN)，循环神经网络(RNN)，并将分类结果与卷积神经网络进行比较(表 2)。其中逻辑回归是在正负两类样本找到一个线性分类边界来划分两类样本的算法；K 近邻则是通过计算新样本与训练集中样本的“距离”来进行新样本的划分，找出训练集中离新样本“距离”最近的 K 个点，K 个点中正(负)类样本更多，则新样本就被预测为正(负)类样本，这里我们使用欧式距离；支持向量机通过寻找离分类边界最近的训练样本点来找到划分超平面，这些样本点被称为支持向量；随机森林是多棵决策树的集成，通过可放回采样，随机选取样本、特征来构造多棵决策树，根据每个树的分类结果来投票共同决定新样本的分类结果；ANN 是最普通神经网络，多层的神经网络通过多次特征的线性组合与非线性函数的激活来得到强大的学习能力；RNN 是 ANN 的一种推广，它使得每一层的神经元之间可以互相连接，从而增加了信息的流动性，往往应用在自然语言处理当中。

表 1　卷积神经网络架构

层数	层类型	卷积核(神经元)个数	卷积核大小	步长	滑动窗口大小
1	卷积	16	5×1	[1, 1]	/
2	最大池化	/	/	[3, 1]	[4, 1]
3	卷积	32	3×1	[1, 1]	/
4	最大池化	/	/	[4, 2]	[3, 2]
5	卷积	64	3×1	[1, 1]	/
6	平均池化	/	/	[1, 1]	[2, 1]
7	全连接	2	/	/	/

注：通过三次卷积操作提取了数据的特征，将数据矩阵的大小进行压缩，但是数据的深度加深，每次池化操作之后都使用了 relu 函数对数据矩阵进行逐元素激活，最后加上一个全连接层将每一个样本进行分类。

5 结果

5.1 ERPs 结果

情绪 Stroop 的 ERP 结果见图 5，方差分析结果表明 P1, P2, N2, P3 和 LPP 这 5 种 ERP 成分对高、低考试焦虑者具有鉴别能力，即在 Fz, FCz, Cz, CPz 和 Pz 点上均有显著结果。具体表现为(以 Cz 点结果为例)，在 5 种成分上，条件主效应在 P2, N2, 和 P3 成分上显著(P2: $F(1, 80) = 9.25$, $p = 0.003$, $\eta^2 = 0.10$; N2: $F(1, 80) = 19.51$, $p < 0.001$, $\eta^2 = 0.20$; P3: $F(1, 80) = 27.86$, $p < 0.001$, $\eta^2 = 0.26$)，在 P1 和 LPP 成分上不显著(Fs (1, 80) < 1.06, ps > 0.307)，组别主效应均不显著(Fs (1, 80) < 1.52, ps > 0.221)，组别与条件交互效应均显著(P1: $F(1, 80) = 11.68$, $p < 0.001$, $\eta^2 = 0.13$; P2: $F(1, 80) = 14.10$, $p < 0.001$, $\eta^2 = 0.15$; N2: $F(1, 80) = 28.55$, $p < 0.001$, $\eta^2 = 0.26$; P3: $F(1, 80) = 22.41$, $p < 0.001$, $\eta^2 = 0.22$; LPP: $F(1, 80) = 16.92$, $p < 0.001$, $\eta^2 = 0.18$)；进一步简单分析表明，高考试焦虑组在考试相关威胁词条件下的 ERP 波幅显著强于中性词条件下(P1: $F(1, 80) = 16.19$, $p < 0.001$, $\eta^2 = 0.17$; P2: $F(1, 80) = 37.88$, $p < 0.001$, $\eta^2 = 0.32$; N2: $F(1, 80) = 78.12$, $p < 0.001$, $\eta^2 = 0.49$; P3: $F(1, 80) = 82.18$, $p < 0.001$, $\eta^2 = 0.51$; LPP: $F(1, 80) = 19.55$, $p < 0.001$, $\eta^2 = 0.20$)，而低考试焦虑组在两种词汇条件下的 ERP 波幅没有显著差异(P1, P2, N2, P3: Fs (1, 80) < 2.06, ps > 0.155; LPP: $F(1, 80) = 4.02$, $p = 0.048$, $\eta^2 = 0.05$，边缘显著)。

5.2 机器学习结果

不同机器学习算法比较的结果见表 2。由于这一批数据正反两类的数目并不均衡，这里我们使用在测试集上的准确率和 F1-score 来评价模型的优劣，F1-score 是样本类别不均衡下一种衡量模型好坏的评价指标，它是基于查准率与查全率的调和平均来定义的，在这一实验中，高考试焦虑人群的数量远多于低考试焦虑人群，因此在高考试焦虑人群上的准确性可能会掩盖低考试焦虑的部分，相对于单一的准确性而言 F1-score 更加全面的衡量了模型在高、低考试焦虑这两类人群上的准确性。通过对不同模型间的各类重要指标进行比较(表 2)，我们发现 CNN 在这一分类任务上的各个重要指标都显著高于其他算法。例如，宿云、胡斌、徐立新、张晓炜和陈婧(2015)在研究中提到的用随机森林对 EGG 信号进行分类的方法，虽然随机森林构建更快，需要调整的参数也更少，但是它在某些噪音较大的分类问题上容易过拟合，且偏向于划分取值较多的特征，因此在当前数据上表现不佳，同时也有研究指出，神经网络往往比随机森林得到的结果更优一些(Strier & Shechter, 2016)。因此我们认为，在对于脑电信号的处理方面，卷积神经网络确实有独特的优势。

表 2 不同机器学习模型的结果对比

机器学习模型	准确率	查准率	查全率	F1-score
卷积神经网络(CNN)	86.5%	84.0%	100%	0.911
逻辑回归(Logistic Regression)	80.3%	83.6%	91.4%	0.868
K 近邻(KNN)	71.8%	71.3%	100.0%	0.817
支持向量机(SVM)	79.0%	78.6%	96.4%	0.865
随机森林(Random Forest)	73.1%	78.7%	84.2%	0.814
人工神经网络(ANN)	82.7%	84.6%	92.9%	0.882
循环神经网络(RNN)	79.2%	77.0%	100%	0.870

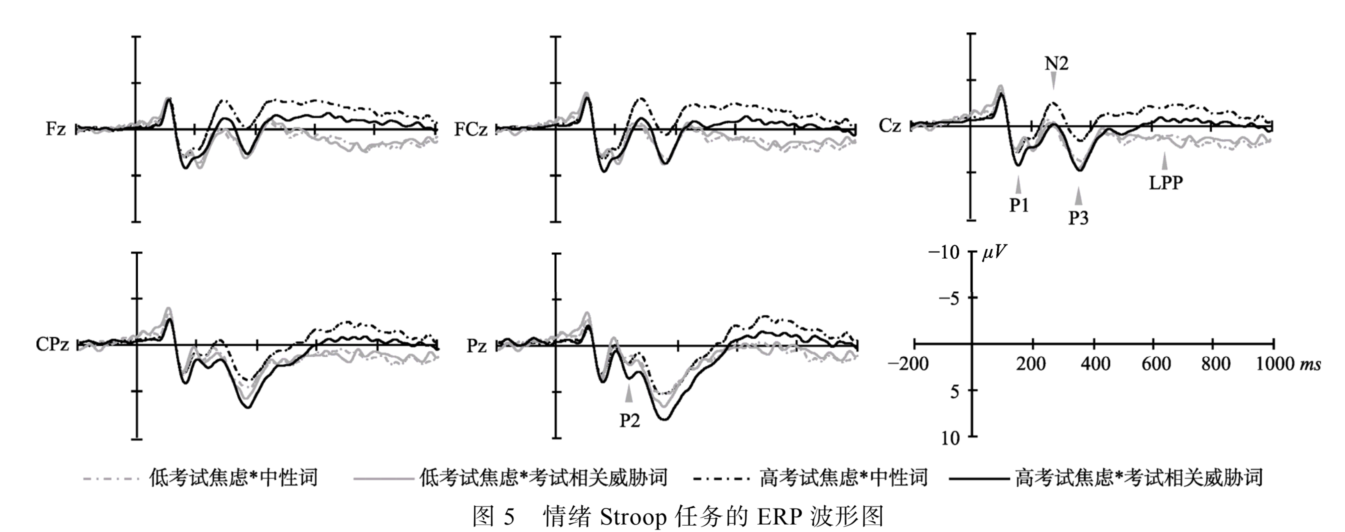


图 5 情绪 Stroop 任务的 ERP 波形图

注：情绪 Stroop 任务中高、低考试焦虑在两种条件(中性词和考试相关威胁词)下的 ERP 总平均波形图(以 Fz, FCz, Cz, CPz 和 Pz 电极点为例)。

6 讨论

本研究提出了一种用于考试焦虑程度的诊断方法。该方法使用 ERPs 技术采集并分析高、低考试焦虑者在情绪 Stroop 中对考试相关威胁词与中性词下的 ERP 成分，采用机器学习(以 CNN 算法为主)建立了一个对被试焦虑与否的分类模型，并且通过一些量化的指标说明 CNN 在这一分类任务上的表现明显好于其它 5 种算法。

首先，情绪 Stroop 范式结合 ERPs 技术确实可以作为一个有效判断考试焦虑程度的诊断任务。具体表现为以下三点：

(1)范式对考试焦虑程度评估的可靠性与有效性。通过比较不同算法在两个指标上的得分(见表 2)可以得出，采用各模型对被试的 ERP 数据进行计算均可以得到较高的准确率和 F1-score，准确率高表示模型对高、低考试焦虑两类人群总的分类能力强，F1-score 高代表模型对于识别高、低考试焦虑人的能力都强并且不会因为两类样本的数量不均匀使模型产生偏差。前人研究中表明当准确率大于 75%，F1-score 大于 0.8，模型就有良好的性能(Demšar, 2006)。本研究中所采用的机器学习算法的结果基本都能满足这一条件，这表明对个体应用情绪 Stroop 范式所采集到的 ERP 数据对个体考试焦虑程度的判断具有稳定性，因此，证明此任务具有良好的效度。

(2)机器学习特征选取的有效性和高度客观性。本研究在机器学习的特征选取中具有重要心理学意义的特征是任务条件(即情绪 Stroop 任务中的考试相关威胁条件与中性条件)和 ERP 成分(即 P1, P2, N2, P3 和 LPP 成分)，而通过对脑电结果的方差分析可以得出这两种重要特征的选取兼具有效性和客观性。首先，方差分析的结果表明这两种特征可以有效区分高、低考试焦虑者。方差结果表明低考试焦虑者在情绪 Stroop 任务中的两种条件下的 ERP 波幅没有显著差异，而高考试焦虑者的 ERP 波幅有显著差异，并体现在各个 ERP 成分上。这说明高考试焦虑者会因为考试相关威胁词的出现而受到干扰，说明结合这两类特征可以有效反映高考试焦虑者认知中视考试相关威胁词为威胁的认知模式(Gootjes, Coppens, Zwaan, Franken, & van Strien, 2011)。其次，在这 5 种 ERP 成分各自代表了不同的心理学意义并且 ERP 波幅可以量化，从而达到诊断的客观性与准确性。具体而言，P1-P2-N2 成分反映的是个体对刺激自动化的注意偏向(Berggren & Derakshan, 2013; Derakshan, Smyth, & Eysenck, 2009)，P3 成分在此类任务中反映的主要是在个体对信息进一步的精细加工(Jo, Schmidt, Inacker, Markowiak, & Hinterberger, 2016; Peng, Cai, & Zhou, 2015)，LPP 成分反映的则是个体对刺激的情绪性自动反应(Cosme & Wiens, 2015; Gootjes et al., 2011)。本研究的 ERP 结果说明当考试相关威胁信息出现时，高考试焦虑者能够迅速注意到并自动化加工这类信息(P1-P2-N2 成分)，之后高考试焦虑者对这类信息的干扰抑制失败，从而对此类信息进行进一步精细加工(P3 成分)，这种对威胁信息的加工随后激发了相应的负性情绪(LPP 成分)，因此本研究中的 ERP 成分可以有效反映高考试焦虑者对考试信息的不合理认知(Chen & Zhou, 2010)。最后，由于高考试焦虑者对考试相关威胁的反应包含高度自动化的反应，他们在进行任务的时候往往很难意识到自己对不同类型词汇的反应，无法猜测任务目的(Yiend, 2010)，因此在使用此模型进行考试焦虑筛查时，可以有效避免由于猜测到任务目的而有意识地进行某种倾向的回答(例如掩盖焦虑程度或者夸大焦虑程度)，达到诊断的客观性与准确性。

(3)情绪 Stroop 范式结合 ERPs 技术评估手段的可操作性。完成一次情绪 Stroop 任务只需要 5 分钟，且实验范式规则简单易懂，具有高度有效性与可靠性(van Bockstaele et al., 2014; Verhaak et al., 2004)，因此能够很好地应用到实际诊断。

其次，不同算法的比较可以得出我们建立的卷积神经网络模型具有良好的区别高、低考试焦虑者的能力。具体表现为以下三点：

(1)高准确率。相比于其他算法，CNN 的算法具有最高的准确率(86.5%)和 F1-score (0.911)。因为 CNN 模型拥有对数据的平移不变性，并且能够捕捉数据的局部特性和提取更高级的特征(Boureau et al., 2010)，因此使得卷积神经网络相对于其他模型而言，对脑电数据的分析具有更高的适用性，因此有比较明显的提升。因此，本研究建立的 CNN 诊断模型具有高度准确性，可靠性和普适性，同时由于 F1-score 很高，这一模型在识别高、低考试焦虑上都有很高的准确度。

(2)诊断精确性。在应用 CNN 模型对考试焦虑进行诊断时，不仅可以对个体是否是考试焦虑者进行诊断，还可以分析出其考试焦虑的程度。在操作层面上，卷积神经网络最后一层输出的是直接的分

类结果，对于一个新的被试，只要输入数据，就可以判定他是否是考试焦虑，而倒数第二层输出的结果是该被试是高考试焦虑或低考试焦虑的概率。这个概率可以反映被试个体的考试焦虑程度，即属于高考试焦虑这一类的概率越大，被试的考试焦虑程度越大。因此，通过对高、低考试焦虑者在情绪 Stroop 中的 ERP 脑电信号进行机器学习的结果可以有效对个体的考试焦虑程度进行客观诊断。

(3)可操作性。虽然对比于传统机器学习方法，卷积神经网络的模型搭建需要仔细的调参，花费更多地时间，但是一旦模型建立，进行预测就会非常快，特别是对于大量数据而言，深度学习模型有非常大的优势。

本研究的局限主要在于两点：首先，本研究中数据量不高，深度学习是数据驱动的模型，即深度模型强大的泛化能力来自于庞大的数据量，由于我们的数据量有限，因此这可能会降低模型的泛化能力。未来可考虑建立大数据数据库，并使用数据增强等一系列手段提升数据量，模型的表现可能有更进一步的提升；其次，本研究提出的综合诊断方法需要借助脑电设备，相对于单纯采用问卷进行诊断还是限制更多。不过随着便携脑电设备的不断发展，此诊断方法会变得越加便利。

在本研究中，我们试图通过卷积神经网络来对考试焦虑进行更加客观的诊断，目标是达到对考试焦虑的及早诊断考试焦虑程度评估。从机器学习的两个重要指标上的表现来看，各类模型是相当有效的，其中，CNN 模型是最适用于 ERP 数据的深度学习，对考试焦虑的诊断及程度判断具有很高的准确率与可靠性。

参 考 文 献

Albert, J., López-Martín, S., & Carretié, L. (2010). Emotional context modulates response inhibition: Neural and behavioral data. *NeuroImage*, *49*(1), 914–921.

Berggren, N., & Derakshan, N. (2013). Attentional control deficits in trait anxiety: why you see them and why you don't. *Biological Psychology*, *92*(3), 440–446.

Boshra, R., Ruiter, K., Reilly, J., & Connolly, J. (2016). Machine learning based framework for EEG/ERP analysis. *International Journal of Psychophysiology*, *108*, 105.

Boureau, Y-L., Bach, F., LeCun, Y., & Ponce, J. (2010, June). Learning mid-level features for recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2559–2566). IEEE.

Cecotti, H., Rivet, B., Congedo, M., Jutten, C., Bertrand, O., Maby, E., & Mattout, J. (2011). A robust sensor-selection method for P300 brain-computer interfaces. *Journal of Neural Engineering*, *8*(1), 016001.

Chen, R., Liu, X. N., & Zhou, R. L. (2011). The attentional bias to threat stimuli in test-anxious students. *Journal of Psychological Science*, *34*(1), 151–154.

[陈睿, 刘潇楠, 周仁来. (2011). 不同程度考试焦虑个体对威胁性刺激注意机制的差异. *心理科学*, *34*(1), 151–154.]

Chen, R., & Zhou, R. (2010). Attentional disengage from test-related pictures in test-anxious students: Evidence from event-related potentials. *International Conference on Brain Informatics*, *6334*, 232–239.

Chen, Z. Y. (2002). Fear of negative evaluation and test anxiety in middle school students. *Chinese Mental Health Journal*, *16*(12), 855–857.

[陈祉妍. (2002). 中学生负面评价恐惧与考试焦虑的相关性. *中国心理卫生杂志*, *16*(12), 855–857.]

Cosme, D., & Wiens, S. (2015). Self-reported trait mindfulness and affective reactivity: A motivational approach using multiple psychophysiological measurese. *PLoS ONE*, *10*(3), e0119466.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dennis, T. A., & Chen, C-C. (2009). Trait anxiety and conflict monitoring following threat: An ERP study. *Psychophysiology*, *46*(1), 122–131.

Derakshan, N., Smyth, S., & Eysenck, M. W. (2009). Effects of state anxiety on performance using a task-switching paradigm: An investigation of attentional control theory. *Psychonomic Bulletin & Review*, *16*(6), 1112–1117.

Diegomantecón, J. (2015). Instrument adaptation in cross-cultural studies of students' mathematics-related beliefs: Learning from healthcare research. *Compare: A Journal of Comparative and International Education*, *45*(4), 545–567.

Donaldson, K. R., Ait Oumeziane, B., Hélie, S., & Foti, D. (2016). The temporal dynamics of reversal learning: P3 amplitude predicts valence-specific behavioral adjustment. *Physiology and Behavior*, *161*, 24–32.

Edwards, M. S., Burt, J. S., & Lipp, O. V. (2010). Selective attention for masked and unmasked threatening words in anxiety: Effects of trait anxiety, state anxiety and awareness. *Behaviour Research and Therapy*, *48*(3), 210–218.

Felmingham, K. L., Stewart, L. F., Kemp, A. H., & Carr, A. R. (2016). The impact of high trait social anxiety on neural processing of facial emotion expressions in females. *Biological Psychology*, *117*, 179–186.

Feng, X. (2003). Result representation and method application: Analysis of 141 investigations. *Sociological Research*, *18*(2), 28–38.

[风笑天. (2003). 结果呈现与方法运用——141 项调查研究的解析. *社会学研究*, *18*(2), 28–38.]

Fotin, S. V., Yin, Y., Haldankar, H., Hoffmeister, J. W., & Periaswamy, S. (2016, March). Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In *Medical Imaging 2016: Computer-Aided Diagnosis* (Vol. 9785, p. 97850X). International Society for Optics and Photonics.

Gootjes, L., Coppens, L. C., Zwaan, R. A., Franken, I. H. A., & van Strien, J. W. (2011). Effects of recent word exposure on emotion-word Stroop interference: An ERP study. *International Journal of Psychophysiology*, *79*(3), 356–363.

Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y-J. (2011). Beyond valence and magnitude: A flexible evaluative coding system in the brain. *Neuropsychologia*, *49*(14), 3891–3897.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Jo, H-G, Schmidt, S., Inacker, E., Markowiak, M., & Hinterberger, T.

(2016). Meditation and attention: a controlled study on long-term meditators in behavioral performance and event-related potentials of attentional control. *International Journal of Psychophysiology*, *99*, 33–39.

Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, *30*(1), 197–215.

Kalanthroff, E., Henik, A., Derakshan, N., & Usher, M. (2016). Anxiety, emotional distraction, and attentional control in the Stroop task. *Emotion*, *16*(3), 293–300.

Kanske, P., & Kotz, S. A. (2012). Effortful control, depression, and anxiety correlate with the influence of emotion on executive attentional control. *Biological psychology*, *91*(1), 88–95.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., … Com, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. *In International conference on machine learning* (pp. 1378–1387).

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. *The Handbook of Brain Theory and Neural Networks* (pp. 255–258). M. A. Arbib, Ed.Cambridge, MA: MIT Press.

Lee, A. (2015). Comparing deep neural networks and traditional vision algorithms in mobile robotics. *Swarthmore University*. Retrieved from http://cs.swarthmore.edu

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*. Retrieved from https://arxiv.org/abs

Lotz, C., & Sparfeldt, J. R. (2017). Does test anxiety increase as the exam draws near? – students' state test anxiety recorded over the course of one semester. *Personality and Individual Differences*, *104*, 397–400.

Lowe, P. A., Lee, S. W., Witteborg, K. M., Prichard, K. W., Luhr, M. E., Cullinan, C. M., ... Janik, M. (2008). The Test Anxiety Inventory for Children and Adolescents (TAICA) examination of the psychometric properties of a new multidimensional measure of test anxiety among elementary and secondary school students. *Journal of Psychoeducational Assessment*, *26*(3), 215–230.

Lu, Y., Jiang, H., & Liu, W. (2017, September). Classification of EEG signal by STFT-CNN framework: identification of right-/left-hand motor imagination in BCI systems. In *The 7th International Conference on Computer Engineering and Networks* (Vol. 299, p. 001).

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188–5196).

Mochcovitch, M. D., da Rocha Freire, R. C., Garcia, R. F., & Nardi, A. E. (2014). A systematic review of fMRI studies in generalized anxiety disorder: evaluating its neural and cognitive basis. *Journal of Affective Disorders*, *167*, 336–342.

Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science*, *44*(6), 567–581.

Morel, S., George, N., Foucher, A., Chammat, M., & Dubal, S. (2014). ERP evidence for an early emotional bias towards happy faces in trait anxiety. *Biological Psychology*, *99*(1), 183–192.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814).

Newman, E. (1996). *No more test anxiety: effective steps for taking tests and achieving better grades.* Los Angeles: Learning Skills Publications.

Peng, M., Cai, M., & Zhou, R. (2015). Processing of task-irrelevant emotional faces impacted by implicit sequence learning. *NeuroReport*, *26*(17), 1056–1060.

Putwain, D. W., Langdale, H. C., Woods, K. A., & Nicholson, L. J. (2011). Developing and piloting a dot-probe measure of attentional bias for test anxiety. *Learning and Individual Differences*, *21*(4), 478–482.

Raz, S., Dan, O., Arad, H., & Zysberg, L. (2013). Behavioral and neural correlates of emotional intelligence: An event-related potentials (ERP) study. *Brain Research*, *1526*, 44–53.

Righi, S., Mecacci, L., & Viggiano, M. P. (2009). Anxiety, cognitive self-evaluation and performance: ERP correlates. *Journal of Anxiety Disorders*, *23*(8), 1132–1138.

Sarason, I. G. (1978). The test anxiety scale: Concept and research. In *Stress and Anxiety* (Vol. 5, pp. 193–216). Washington DC: Hemisphere.

Schutz, P. A., Davis, H. A., & Schwanenflugel, P. J. (2002). Organization of concepts relevant to emotions and their regulation during test taking. *The Journal of Experimental Education*, *70*(4), 316–342.

Seijdel, N., Ramakrishnan, K., Losch, M., & Scholte, S. (2016). Overlap in performance of CNN's, human behavior and EEG classification. *Journal of Vision*, *16*(12), 501.

Strier, R., & Shechter, D. (2016). Visualizing access: Knowledge development in university-community partnerships. *Higher Education*, *71*(3), 343–359.

Su, Y., Hu, B., Xu, L. X., Zhang, X. W., & Chen, J. (2015). EEG-data-oriented knowledge modeling and emotion recognition. *Chinese Science Bulletin (Chinese Version)*, *60*(11), 1002–1009. https://doi.org/10.1360/N972014-00829

[宿云，胡斌，徐立新，张晓炜，陈婧. (2015). 面向脑电数据的知识建模和情感识别. *科学通报*, *60*(11), 1002–1009.]

Thomas, S. J., Johnstone, S. J., & Gonsalvez, C. J. (2007). Event-related potentials during an emotional Stroop task. *International Journal of Psychophysiology*, *63*(3), 221–231.

Tillman, C. M., & Wiens, S. (2011). Behavioral and ERP indices of response conflict in Stroop and flanker tasks. *Psychophysiology*, *48*(10), 1405–1411.

van Bockstaele, B., Verschuere, B., Tibboel, H., de Houwer, J., Crombez, G., & Koster, E. H. W. (2014). A review of current evidence for the causal impact of attentional bias on fear and anxiety. *Psychological Bulletin*, *140*(3), 682–721.

Verhaak, C. M., Smeenk, J. M., van Minnen, A., & Kraaimaat, F. W. (2004). Neuroticism, preattentive and attentional biases towards threat, and anxiety before and after a severe stressor: A prospective study. *Personality and Individual Differences*, *36*(4), 767–778.

Wabnitz, P., Martens, U., & Neuner, F. (2016). Written threat: electrophysiological evidence for an attention bias to affective words in social anxiety disorder. *Cognition and Emotion*, *30*(3), 516–538.

Wang, C. K. (2001). Reliability and validity of test anxiety scale-Chinese version. *Chinese Mental Health Journal*,

15(2), 96–97.
[王才康. (2001). 考试焦虑量表在大学生中的测试报告. *中国心理卫生杂志*, *15*(2), 96–97.]
Wang, W.-W., Xie, X., & Shao, F. (2008). Early-onset depression and its neural basis. *Advances in Psychological Science*, *16*(3), 411–417.
[王玮文, 谢希, 邵枫. (2008). 早发性抑郁及其神经基础. *心理科学进展*, *16*(3), 411–417.]
Wang, Y-N., & Sun, B-Y. (2017). Cigarette craving EEG classification based on convolution neural networks. *Computer Systems & Applications*, *26*(6), 256–260.
[王艳娜, 孙丙宇. (2017). 基于卷积神经网络的烟瘾渴求脑电分类. *计算机系统应用*, *26*(6), 256–260.]
Yan, H., & Lu, L. (2014). Effects of exam stress on psychosomatic response saliva immuneglobulin and cortisol among medical college student. *Chinese Journal of School Health*, *35*(6), 813–816.
[闫慧, 卢莉. (2014). 考试应激对医学生心身反应唾液免疫球蛋白及皮质醇的影响. *中国学校卫生*, *35*(6), 813–816.]
Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.
Yiend, J. (2010). The effects of emotion on attention: A review of attentional processing of emotional information. *Cognition and Emotion*, *24*(1), 3–47.
Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*. Retrieved from https://arxiv.org/abs
Zeidner, M., & Matthews, G. (2005). Evaluation anxiety. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook ofcompetence and motivation* (pp. 141–163). London: Guildford Press.
Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer, Cham.

Classification of test-anxious individuals using Event-Related Potentials (ERPs): The effectiveness of machine learning algorithms

ZHANG Wenpei[1,2]; SHEN Qunlun[3]; SONG Jintao[1]; ZHOU Renlai[1]

([1] *Department of Psychology, Nanjing University, Nanjing*, 210023, *China*)

([2] *Department of Business Administration, School of Business, Anhui University of Technology, Maanshan*, 243032, *China*)

([3] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences*, 100190, *China*)

Abstract

Individuals with test anxiety always treat tests/examinations as a potential threat. This cognitive mode impairs these individuals' cognition, attention and emotions. A traditional method classifying subjects either as high or low on test anxiety (i.e., HTA or LTA, respectively) relies on questionnaire data. Questionnaire data may be unstable due to the subjective nature of participants' attitudes, implying a reduced classification accuracy. In search for higher levels of (data) stability and classification accuracy a new classification approach is proposed. This new approach overcomes subjective data's negative impact on classification accuracy by relying on event-related potential (EPR) data (also referred to as ERPs), objective (multivariate, longitudinal) data which adequately capture participants' reactions to relevant stimuli (over time). However, as ERP data may still be somewhat unstable due to individual differences between participants, (machine) learning algorithms are adopted as their 'learning' feature may increase both the stability of ERP data and classification accuracy.

This study recruited 57 HTA participants and 25 LTA participants based on: (a) Test Anxiety Scale (TAS) scores, and (b) (two) specialists' psychological diagnostic results on a single participant. Reliance on the emotional Stroop (ES) paradigm in combination with ERP technology enabled the assessment of participants' cognitive mode related to test anxiety. In ES, the information on the ERP components P1, P2, N2, P3 and LPP ERP were selected as input for seven commonly used machine learning algorithms: Convolutional Neural Network (CNN), Logistic Regression (LR), K Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), and Recurrent Neural Network (RNN). To compare the classification accuracy of these algorithms (using the complete sample of HTA and LTA subjects) important indexes (i.e., accuracy and F1-score) were calculated and compared across these algorithms.

The results showed that: (a) the ERPs data collected in ES allow effective differentiation between HTA and LTA (P1: $F(1, 80) = 11.68$, $p < 0.001$, $\eta^2 = 0.13$; P2: $F(1, 80) = 14.10$, $p < 0.001$, $\eta^2 = 0.15$; N2: $F(1, 80) = 28.55$, $p < 0.001$, $\eta^2 = 0.26$; P3: $F(1, 80) = 22.41$, $p < 0.001$, $\eta^2 = 0.22$; LPP: $F(1, 80) = 16.92$, $p < 0.001$, $\eta^2 = 0.18$); (b)

classification on the basis of ERP data using machine learning algorithms shows high accuracy and stability, that is the classification accuracy of all seven algorithms is found to be high as evidenced by an accuracy index of 71.8% or higher (CNN: 86.5%, LR: 80.3%, KNN: 71.8%, SVM: 79.0%, RF: 73.1%, ANN: 82.7%, and RNN: 79.2%) and an F1-score of 0.814 or higher (CNN: 0.911, LR: 0.868, KNN: 0.817, SVM: 0.865, RF: 0.814, ANN: 0.882, and RNN: 0.870); (c) CNN outperforms the other six common machine learning algorithms showing both the highest accuracy index and F1-score. Moreover, as over and above this (relative) superiority CNN combines the (technical) property known as 'shift invariance' and robustness to noise, the algorithm may be considered ideal for effectively classifying test anxious individuals using ERP data.

It is concluded that: (a) as manifested by its 'discriminatory' nature and stable classification performance (as evidenced by all machine learning algorithms' favorable values for all important indices) reliance on the ES paradigm enables machine learning leading up to effective diagnosis of test anxiety; and (b) participants' classification into HTA and LTA by relying on ERP data which are subsequently analyzed by means of the machine learning algorithm CNN is (most) effective (i.e., as benchmarked against six other commonly used machine learning algorithms). Consequently, using ES in combination with ERP technology and the CNN machine learning algorithm can be conceived as an ideal method for diagnosing test anxiety.

Key words　machine learning; test anxiety; emotional Stroop; ERPs